ANALYSIS OF PHONETIC/PROSODIC FEATURES IN INTERACTION STAGES

Daniel Duran[1], *Ronald Böck*[2]
[1]*Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS) Berlin,*
[2]*Otto von Guericke University Magdeburg*
duran@leibniz-zas.de, ronald.boeck@ovgu.de

Abstract: In daily life, any kind of verbal and non-verbal communication can be considered as an interaction. Therefore, the entire lifetime comprises of a sequence of a multitude of interactions, being separated in stages. These stages are usually related to a specific aspect in the communication. The current paper investigates phonetic and prosodic features which can be used to understand differences in interaction stages from verbal cues. In this sense, the LAST MINUTE Corpus, comprising 133 naturalistic human-machine interactions, is analysed, especially considering (consecutive) interaction stages. A collection of phonetic/prosodic features, mainly extracted on utterance-level, were statistically analysed. The results indicated that especially the speech rate, articulation rate, and average syllable rate show significant differences in the observed interaction stages.

1 Introduction

Following the argumentation of Watzlawick et al. [1], any kind of verbal or non-verbal signalling can be interpreted as interaction. This aspect was extended to sequences of particular interactions in [2], where each specific interaction can be concatenated to a collection or sequence of interactions. Therefore, quite philosophically, the entire lifetime can be considered as a chronology of interactions or interaction stages. As defined in [2] a "stage is an interaction period which is related to either a collection of particular topics or an aggregation of actions triggered by a certain event". In this sense, the question arises, how different interaction stages can be (automatically) distinguished, especially from verbal cues. A first attempt to answer the question is presented in [3], investigating appropriate low-level features. These investigations are extended to classification experiments in [4] and are discussed in combination in [2]. These results indicate trends in automatic assessment of interaction stages based on verbal cues. Given these evidences, we extended the analyses, regarding the aspect: Which phonetic and prosodic features can be used as indicators for the distinction of interaction stages? The experiments were based on the LAST MINUTE Corpus (LMC) (cf. Section 2), which also allows a direct comparison to the work presented in [3, 4].

1.1 Research Questions

Aiming on an understanding of differences in phonetic/prosodic low-dimensional features in different interaction stages – in addition to low-level features considered in [3] –, we particularly investigated the following specific research questions:

RQ 1: How does a change of interaction stage (from a relaxed to a stressful situation) affect speaking rate?

RQ 2: How does such a change affect a speaker's vowel space size?

RQ 3: How does such a change affect a speaker's pitch range?

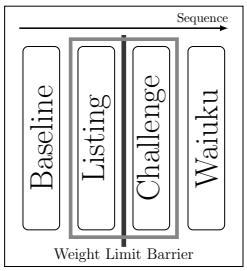

Figure 1 – Stages in the LMC, indicating the observed two consecutive stages (green box) and the utilised barrier.

1.2 Related Work

Regarding related work, we see already some overviews in [2, 3, 4], discussing aspects of interaction stages and feature sets based on low-level descriptors and functionals. Therefore, we just briefly highlight some works according to feature observations and classification of interaction stages.

Considering (optimal) feature sets for automatic classification acoustics in interactions, we see that even for different purposes (affect detection, speech processing, or addressee detection) "just" a few sets seem to be commonly used in the community. These are mainly derived from low-level descriptors applying functionals to achieve a range of 39 basic features to 988 extended features up to 6,373 advanced features (cf. [2, 5]). However, in terms of changes in interactions no common sets have be established, yet. Therefore, [3] investigate this aspect using a statistical approach deriving relevant features based on the *emobase* feature set (cf. [6]). Although, those feature sets already show a relation to prosodic features, no specific analysis was provided. Therefore, in the current paper, we contribute to this discussion, extending the perspective to more linguistically motivated phonetic/prosodic analyses as presented in Section 3. This analyses can be directly compared to the work of [3] since both investigations use the same corpus (cf. Section 2).

Furthermore, in [4] consecutive interaction stages are classified based on features derived in [3]. As further argued in [2], the detection of changes in interaction stages is related to aspects of topic changes. Aminikhanghahi and Cook [7] provide an overview on this issue, also providing a definition on topic changes. Particularly, in linguistics, the aspect of topic changes is already considered for a longer time (cf. e.g. [8]). Further, in [9] reflects this aspect and also provide an overview of the beginnings in an automatic processing of changes in topics. However, since our focus is on feature investigations, we refer to the mentioned literature for further reading and rather focus on related work on phonetic, prosodic, and intonational analyses.

How the situational context affects speech production has been a research topic addressed in numerous studies. Speakers adapt their speech, among other things, on the phonetic and prosodic levels to their human interlocutors [10, 11] and even in Human-Machine Interaction (HMI) [12] based on the social and situated context of the interaction as well as individual psychological and cognitive factors (which is important, for example, in automotive applications of HMI [13]). Under certain types of cognitive load, for example, articulation rate has been found in one study to decrease, as well as the pitch range and the vowel space to be more centralised [14, 15]. Other studies suggest that formant shifts may be vowel specific, showing more complex patterns in relation to cognitive load [16], or that pitch range may decrease and the speaking rate increases while the vowel space shows no effect of cognitive workload [17]. Examples like these show that the reported results in the literature are very specific to certain situations and, thus in general, still not conclusive.

Table 1 – Examples of interactions in the two interaction stages. The table is adapted from [3], whereas the wizard's text (W) was taken verbatim from [18].

	Listing stage		
	German		English
W:	*Sie können jetzt aus der Rubrik Hosen und Röcke auswählen*	W:	*You may now choose from the category trousers and skirts*
U:	*Zwei Jeans*	U:	*Two jeans trousers*
W:	*Zwei Jeans wurden hinzugefügt*	W:	*Two jeans trousers have been added*
	Challenging stage		
	German		English
W:	*Der Artikel Kosmetikset kann nicht hinzugefügt werden. Anderenfalls würde die maximale Gewichtsgrenze des Koffers überschritten werden*	W:	*The item cosmetic set cannot be added. Otherwise the weight limit of your suitcase will be exceeded*
U:	*Dann... einen ((leise)) ... ((klopft)) ... dann bitte einen Anorak raus*	U:	*Then... a... ((quietly)) ... ((knocks)) ... then please take out an anorak*

2 LAST MINUTE Corpus

The investigations of the paper were based on the LMC [18, 19] comprising a naturalistic, naïve HMI in the sense of Valli [20]. This corpus was used to allow a direct comparison to the work of [3, 4].

As summarised in [2], the corpus' story "revolves around a journey to a currently unknown place called »Waiuku«. During the luggage preparation the participant receives more and more information on the journey's location and the current weather conditions." Each recording session takes roughly 30 minutes, where the plot can be divided into six phases reflecting particular issues presented to the participants, being accompanied by specific system reactions. The corpus's design is based on the theoretical aspects of Halliday's model of "systemic functional grammar" (cf. [18]). Some phases can be combined, so that the experimental setting divides the interaction into four distinct stages (cf. Figure 1), where each stage represents an increasing task difficulty [18]. Every stage is marked by a so-called barrier [21], allowing a dedicated alignment of the participant's utterances. In the current study, we focussed on the *Weight Limit Barrier* [21] separating the *Listing* and *Challenge* stage (cf. Figure 1; for further details cf. [2, 18, 19]). The two stages are characterised as follows:

- In the *Listing* stage the participants collect items in a command-like fashion, being interupted by the system which lists the current content of the luggage. Therefore, a relaxed but also somehow bored state is assumed for the participant.
- In contrast, in the *Challenge* stage the participant has to rearrange (for the first time) the luggage's content, since the airlines weight limit is reached. This results in a more stressful situation for the participant.

In general, the LMC comprises multimodal interactions of 133 participants, where in the current study 89 sessions (48 female, 41 male, 43 younger than 30 years, 46 older than 60 years) were analysed; providing suitable material. The analyses were conducted on utterance-level per participant as in relation to [4]. Examples of interactions are given in Table 1. For further prototypical interactions we refer to [3, 21].

3 Methods

The speech data in the LMC is not (yet) manually segmented and annotated at the *phonetic level*. We therefore applied the following approach based on an automatic method to detect syllable nuclei [22]. Each user's utterance is individually analysed using PRAAT [23]. First, based on the automatically segmented syllables, the following measures were extracted: the speech rate (i.e. the number of syllables divided by duration), the articulation rate (number of syllables divided by phonation time), and the average syllable duration (ASD; the speaking time per syllable).

In order to measure the vowel space size, we extracted the first three formants F1, F2, and F3 in Bark with a custom PRAAT script using the Burg algorithm at each estimated syllable nucleus. The vowel space size is measured by the average vowel dispersion, i.e. the EUCLIDEAN distance of each vowel token from the vowel space centre within the acoustic F1$\times$F2$\times$F3 feature space, excluding statistical outliers. Features related to pitch (minimum, maximum, standard deviation, etc.) have been extracted from PRAAT's Voice report across each utterance, excluding silent pauses.

All phonetic and prosodic data has been statistically modelled with Linear-Mixed Effects Regression (LMER) in R, using the *lme4* and *lmerTest* packages [24, 25]. We fit separate linear mixed-effects regression models for *speech rate*, *articulation rate*, *ASD*, *vowel dispersion*, and *pitch range* as dependent variables. As fixed effects, we included *stage* (before vs after the barrier), *gender* (M=male vs F=female), and age group (Y=younger vs O=older) as fixed effects, including two-way interactions, and speaker ID as random effect.

4 Results

The presented analyses contributed to the aspect of acoustic-based verbal differentiation of interaction stages. In [3], the authors investigated low-level descriptors extracted on utterance level utilising openSMILE. We extended such analyses to prosody-based features. The assessment revealed that speech rate, articulation rate, and ASD are significantly higher in the *Challenge* stage (cf. Figure 1). Especially, speech rate increased highly significantly in the second stage. Regarding the other features related to speech rate, significant differences can be found. Similar trends were also found in several low-level descriptors (details can be found in [2, 3]).

In the following, we extend our discussion: In total 13,721 "segments" (i.e. estimated syllable nuclei) have been acoustically analysed and submitted to the regression analysis. Table 2 shows a summary of the estimates of the fixed effect coefficients. The regression model for *speech rate* as the dependent variable (first column) shows a weakly significant effect for *stage* ($p = 0.011$, t-test with Satterthwaite's method as implemented in the *lmerTest* R package [25]). After the barrier, the speech rate increased. Older speakers, however, show an almost reversed pattern ($p = 0.0145$), as indicated by the significant interaction between *stage* and *age*. The increase in speech rate can thus be attested only for the "younger" group of speakers (younger than 30 years).

Other preliminary results on *articulation rate* or *ASD* show a similar trend, albeit the effects are only weak: articulation rate is larger in the second stage after the barrier and ASD is smaller. For *vowel dispersion* on all vowel tokens, we found no significant effects of stage. There is a tendency for a more compact vowel space in the *Challenge* stage after the barrier, i.e. a smaller vowel dispersion. Since different vowels (e.g. front vs back vowels) may be affected differently or to different degrees by the change in (especially consecutive) interaction stages, future work will need to tease apart different vowel types in order to identify possible patterns of formant shifts due to interaction stages.

Table 2 – LMER fixed effects' estimates for the dependent variables *speech rate*, *articulation rate*, *ASD*, *dispersion*, and *pitch range*. Significance levels are marked as: *** $p < 0.001$, * $p < 0.05$, † $p < 0.1$.

	speech rate	articulation rate	ASD	dispersion	pitch range
(Intercept)	2.997 ***	3.798 ***	0.287 ***	2.152 ***	1.554 ***
stage [after]	0.348 *	0.194 †	−0.017 †	−0.046	−0.200
gender [F]	0.065	0.056	0.001	0.118	1.099 ***
age [O]	0.137	−0.143	0.013	0.089	0.442 *
stage [after] × gender [F]	0.155	0.032	−0.005	0.087	−0.103
stage [after] × age [O]	−0.382 *	0.027	0.003	−0.078	0.323 *
gender [F] × age [O]	−0.051	0.034	−0.008	0.020	−0.565 *

For *pitch range* (defined as the average short term range in f0 across each utterance), we found significant effects for *gender* (which is expected, as we did not normalise the data in order to remove gender differences) and *age*. Further, significant interactions were revealed between *stage*×*age* and *gender*×*age* such that the pitch range of younger speakers tends to decrease in the second interaction stage, with a significantly larger difference for female speakers than for male speaker.

5 Conclusion

Based on the LMC (cf. Section 2), we analysed prosodic features, extracted on utterance level, with respect to differences in (consecutive) interaction stages. In particular, the stages *Listing* and *Challenge* were investigated (cf. Figure 1), being separated by the *Weight Limit Barrier* [21]. The assessment revealed that speech rate and articulation rate are significantly higher in the *Challenge* stage, whereas ASD decreased. Other features show only a minor effect or no significant difference regarding the interaction stage. Our results on speech rate and vowel dispersion seem to be in line with previous studies (cf. [14]) on speech production under increased cognitive workload. In combination with the low-level descriptors, identified in [3], we could obtain a better understanding of acoustic-phonetic-based and prosody-based assessment of interaction stages and contributed to the discrimination of such stages.

However, at least two directions of research should be considered in future work. At first, the effects of changes in interaction stages on different vowel types need to be observed in more details, especially in order to identify relevant patterns. Further, the issue is related to the question why such formant shifts occur or are influenced by interaction stages. Secondly, the work of [4] should be extended to include the current findings in the automatic classification of (consecutive) interaction stages. This aspect additionally contributes to the ongoing discussion of appropriate feature sets in acoustic based HMI analyses.

Acknowledgements

We acknowledge funding by Vector Stiftung within the project "Der Faktor Mensch in der Mensch-Maschine Interaktion" (`https://vector-stiftung.de`). Further, we also acknowledge support by "Adaptive Strategies in Assistive Technologies in Multi-Person Interaction" (ASAMI) funded by the Federal State of Sachsen-Anhalt, Germany (grant number: I 138).

References

[1] WATZLAWICK, P., J. H. BEAVIN, and D. D. JACKSON: *Menschliche Kommunikation: Formen, Störungen, Paradoxien*. Verlag Hans Huber, Bern, Switzerland, 2007.

[2] BÖCK, R.: *Anticipate the User: Multimodal Analyses in Human-Machine Interaction towards Group Interactions*. TUDpress, Dresden, Germany, 2020.

[3] BÖCK, R., O. EGOROW, and A. WENDEMUTH: *Speaker-group specific acoustic differences in consecutive stages of spoken interaction*. In *Proc. of the 28. Konferenz Elektronische Sprachsignalverarbeitung*, pp. 211–218. TUDpress, 2017.

[4] BÖCK, R., O. EGOROW, and A. WENDEMUTH: *Acoustic detection of consecutive stages of spoken interaction based on speaker-group specific features*. In *Proc. of the 29. Konferenz Elektronische Sprachsignalverarbeitung*, pp. 247–254. TUDpress, 2018.

[5] SCHULLER, B.: *Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends*. *Communications of the ACM*, 61(5), pp. 90–99, 2018.

[6] EYBEN, F., F. WENINGER, F. GROSS, and B. SCHULLER: *Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor*. In *Proc. of the 21st ACM International Conference on Multimedia*, pp. 835–838. ACM, Barcelona, Spain, 2013.

[7] AMINIKHANGHAHI, S. and D. J. COOK: *A survey of methods for time series change point detection*. *Knowledge and Information Systems*, 51(2), pp. 339–367, 2017.

[8] COVELLI, L. H. and S. O. MURRAY: *Accomplishing topic change*. *Anthropological Linguistics*, 22(9), pp. 382–389, 1980.

[9] SHRIBERG, E., A. STOLCKE, D. HAKKANI-TÜR, and G. TÜR: *Prosody-based automatic segmentation of speech into sentences and topics*. *Speech Communication*, 32(1), pp. 127–154, 2000.

[10] LEWANDOWSKI, N. and M. JILKA: *Phonetic Convergence, Language Talent, Personality and Attention*. *Frontiers in Communication*, 4, p. 18, 2019. doi:10.3389/fcomm.2019.00018.

[11] DURAN, D. and N. LEWANDOWSKI: *Cognitive factors in speech production and perception: A socio-cognitive model of phonetic convergence / Kognitive Faktoren in der Sprachproduktion und Perzeption: Ein sozio-kognitives Modell der phonetischen Konvergenz*. In M. MATEŠIĆ and A. MEMIŠEVIĆ (eds.), *Language and Mind: Proceedings from the 32nd International Conference of the Croatian Applied Linguistics Society*, pp. 15–31. Peter Lang, Berlin, 2020.

[12] GESSINGER, I., B. MÖBIUS, S. LE MAGUER, E. RAVEH, and I. STEINER: *Phonetic accommodation in interaction with a virtual language learning tutor: A Wizard-of-Oz study*. *Journal of Phonetics*, 86, p. 101029, 2021. doi:10.1016/j.wocn.2021.101029.

[13] DURAN, D. and N. LEWANDOWSKI: *Untersuchung der kognitiven Beanspruchung durch Sprachassistenzsysteme*. In A. BERTON, U. HAIBER, and W. MINKER (eds.), *Elektronische Sprachsignalverarbeitung 2018: Tagungsband der 29. Konferenz*, pp. 159–166. TUDpress, 2018. URL `http://www.essv.de/paper.php?id=403`.

[14] HUTTUNEN, K. H., H. I. KERÄNEN, R. J. PÄÄKKÖNEN, R. PÄIVIKKI ESKELINEN-RÖNKÄ, and T. K. LEINO: *Effect of cognitive load on articulation rate and formant frequencies during simulator flights*. *The Journal of the Acoustical Society of America*, 129(3), pp. 1580–1593, 2011. doi:10.1121/1.3543948.

[15] HUTTUNEN, K., H. KERÄNEN, E. VÄYRYNEN, R. PÄÄKKÖNEN, and T. LEINO: *Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights*. *Applied Ergonomics*, 42(2), pp. 348–357, 2011. doi:10.1016/j.apergo.2010.08.005.

[16] YAP, T. F., J. EPPS, E. AMBIKAIRAJAH, and E. H. C. CHOI: *An investigation of formant frequencies for cognitive load classification*. In *Proc. Interspeech 2010*, pp. 2022–2025. 2010. doi:10.21437/Interspeech.2010-572.

[17] LIVELY, S. E., D. B. PISONI, W. VAN SUMMERS, and R. H. BERNACKI: *Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences*. *The Journal of the Acoustical Society of America*, 93(5), pp. 2962–2973, 1993. doi:10.1121/1.405815.

[18] FROMMER, J., D. RÖSNER, M. HAASE, J. LANGE, R. FRIESEN, and M. OTTO: *Detection and Avoidance of Failures in Dialogues – Wizard of Oz Experiment Operator's Manual*. Pabst Science Publishers, 2012.

[19] RÖSNER, D., J. FROMMER, R. ANDRICH, R. FRIESEN, M. HAASE, M. KUNZE, J. LANGE, and M. OTTO: *Last minute: a novel corpus to support emotion, sentiment and social signal processing*. In *Proc. of the 8th International Conference on Language Resources and Evaluation*, pp. 82–89. ELRA, Istanbul, Turkey, 2012.

[20] VALLI, A.: *The design of natural interaction*. *Multimedia Tools and Applications*, 38(3), pp. 295–305, 2008.

[21] PRYLIPKO, D., D. RÖSNER, I. SIEGERT, S. GÜNTHER, R. FRIESEN, M. HAASE, B. VLASENKO, and A. WENDEMUTH: *Analysis of significant dialog events in realistic human–computer interaction*. *Journal on Multimodal User Interfaces*, 8(1), pp. 75–86, 2014.

[22] DE JONG, N. H. and T. WEMPE: *Praat script to detect syllable nuclei and measure speech rate automatically*. *Behavior Research Methods*, 41(2), pp. 385–390, 2009. doi:10.3758/BRM.41.2.385.

[23] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer*. 2015. URL `http://www.praat.org/`.

[24] BATES, D., M. MÄCHLER, B. BOLKER, and S. WALKER: *Fitting Linear Mixed-Effects Models Using lme4*. *Journal of Statistical Software*, 67(1), pp. 1–48, 2015. doi:10.18637/jss.v067.i01.

[25] KUZNETSOVA, A., P. B. BROCKHOFF, and R. H. B. CHRISTENSEN: *lmerTest package: Tests in Linear Mixed Effects models*. *Journal of Statistical Software*, 82(13), 2017. doi:10.18637/jss.v082.i13.